Development of data driven metatranscriptomic analysis

Ryo Mameda (Hiroshima University)

Abstract

Complex microbiomes are known as playing a significant role in environment, such as soil, ocean, and animal gut. Gene expression analysis using next-generation sequencing (NGS) is a useful approach to highlight their function. This process *in silico* generally comprises three steps which are reference sequence construction, read mapping, and quantification of gene expression. Due to the high diversity of complex microbiomes, metagenomic contigs are often used as reference sequences in place of complete genomes. In the read mapping step, it is known that utilizing both metagenomic and metatranscriptomic reads enables more comprehensive analyses. However, studies systematically evaluating mapping strategies under such conditions remain limited. In addition, transcripts per million (TPM) is widely used as a normalization method in expression quantification. Nevertheless, the potential impact of highly abundant transcripts (particularly ribosomal RNA (rRNA)) on TPM-based quantification has not been fully characterized.

In this study, we analyzed soil metagenomic and metatranscriptomic datasets which is publicly available from the NCBI Sequence Read Archive (SRA). We first evaluated the performance of three common mapping tools (BWA-MEM, Bowtie2, and Minimap2) for aligning both types of reads to predicted protein-coding sequences derived from metagenomic contigs. Among them, BWA-MEM demonstrated significant performance for this purpose. Further analysis revealed that the predicted coding sequences are frequently contaminated rRNA regions. When *in vitro* rRNA depletion was not performed during RNA-seq library preparation, approximately 90% of metatranscriptomic reads were mapped to rRNA regions, indicating a substantial potential bias in expression quantification. To assess the bias, we examined time-series samples from cultured soil microbiomes and calculated TPM values normalized by gene abundance, with and without rRNA-derived reads. Differential expression analysis using DESeq2 identified significantly variable genes across samples (p < 0.05). Notably, TPM values computed without excluding rRNA tended to underestimate the expression of biologically activated genes compared to those calculated after rRNA exclusion. We hope that these findings will contribute advanced methodologies of gene expression analyses in complex microbiomes.